DOCUMENT RESUME

ED 472 474 HE 035 633

AUTHOR	Tamada, Mike
TITLE	Predictors of Graduation Rates: Why They Can Be Expected To Have Greater Predictive Value at the National Level Than at the Institutional Level. AIR 2002 Forum Paper.
PUB DATE	2002-06-00
NOTE	30p.; Paper presented at the Annual Forum of the Association for Institutional Research (42nd, Toronto, Ontario, Canada, June 2-5, 2002). Originally presented at the Annual Meeting of the California Association for Institutional Research (Sacramento, CA, November 2001).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*College Graduates; Error of Measurement; *Graduation Rate; Higher Education; National Norms; State Norms; *Statistical Analysis
IDENTIFIERS	Between Group Differences; Time Series Analysis; Within Group Differences

ABSTRACT

National studies of students, and studies that compare institutions, have identified many predictors of students' graduation rates, including socioeconomic status and admission selectivity. When there predictive variables are applied to data from an individual school, it may be found that they have less predictive power. This paper presents a theoretical explanation of why this pattern might be expected, and uses the Statistical Analysis System (SAS) to estimate a hierarchical linear model of the size of the interinstitutional and intrainstitutional effects of admission selectivity on graduation rates for a 7-year panel of data on several colleges and universities. This statistical phenomenon can be interpreted in a variety of ways: (1) measuring "within groups" effects versus "between groups" effects; (2) analyzing "time-series" data versus "cross-sectional" data; (3) regression analysis with unobserved or omitted variables that cause the error term to be correlated with the explanatory variables; (4) intrainstitutional versus interinstitutional or national data. In short, these interpretations represent the difference between looking ad data from one college and looking at data from national sources. (Contains 10 figures and 7 references.) (Author/SLD)

PREDICTORS OF GRADUATION RATES: WHY THEY CAN BE EXPECTED TO HAVE GREATER PREDICTIVE VALUE AT THE NATIONAL LEVEL THAN AT THE INSTITUTIONAL LEVEL

Presented at the Association for Institutional Research Forum

Toronto, Canada, June 2002

Originally Presented at the California Association for Institutional Research

Annual Conference, Sacramento, California

November 2001

Mike Tamada

Director of Institutional Research

Occidental College

PREDICTORS OF GRADUATION RATES: WHY THEY CAN BE EXPECTED TO HAVE GREATER PREDICTIVE VALUE AT THE NATIONAL LEVEL THAN AT THE INSTITUTIONAL LEVEL

Abstract:

National studies of students, and studies which compare institutions, have identified many predictors of students' graduation rates including socio-economic status and admission selectivity. When you apply these predictive variables to data from your own school, you may find that they have less predictive power. This paper presents a theoretical explanation of why we might expect to see this pattern, and uses SAS to estimate a hierarchical linear model of the size of the inter-institutional and the intra-institutional effects of admission selectivity on graduation rates for a seven-year panel of data on several colleges and universities.

This statistical phenomenon can be interpreted in a variety of ways: measuring "within groups" effects vs "between groups" effects; analyzing "time-series" data vs "cross-sectional" data; regression analysis with unobserved or omitted variables which cause the error term to be correlated with the explanatory variables; intra-institutional vs inter-institutional or national data – in short, the distinction between looking at data from your own college and data from national data sources.

I. INTRODUCTION

National studies of students, and studies which compare institutions, have identified many predictors of students' graduation rates, including socio-economic status and admission selectivity. When you apply these predictive variables to data from your own school, you may find that they have less predictive power. This paper presents a theoretical explanation of why we might expect to see this pattern, and uses SAS to estimate a hierarchical linear model of the size of the inter-institutional and the intra-institutional effects of admission selectivity on graduation rates for a panel of data on several colleges and universities. One lesson is that there can be a big difference between comparing data from your school to data from other schools, and comparing your school's data across time.

Many of the major ideas in this paper – looking at the effect of first-generation college status on graduation rates, looking at admit rates as a predictor of graduation rates, and using hierarchical linear models – were directly inspired by previous presentations at the California Association for Institutional Research (CAIR) Annual Conference and other conferences. So I hope that yet another lesson of this paper is that conference attendance can lead to a fruitful exchange of ideas.

II. BACKGROUND

At the 1998 CAIR Conference, presenters from Pasadena City College and the University of LaVerne noted that first-generation college status was not a good predictor of retention at their campuses. At my own campus, I had also discovered this to be true.

Yet many national studies have identified socio-economic status, including parental education, as an important predictor of college graduation rates. (Adelman (1999) however has found that parental education and even socio-economic status have weaker or no predictive value when other variables such as "intensity and quality of academic curriculum in high school" are taken into account.)

At another conference in early 2001, a presenter, Steve Butts from Lawrence University in Wisconsin, looked at a predictor of graduation rates which many of us had not examined before: the school's admit rate, that is, what percent of applicants it admitted. He found that when he looked at several schools' admit rates and graduation rates, the correlation was very strong -- higher than .7.

Several of us wondered if a school could similarly look at the admit rates of the cohorts it admitted over the years, and if these admit rates would help predict the cohorts' graduation rates. We went back to our campuses and looked at our data; most, though not all, of us found virtually no relationship.

There is a pattern here: national studies, or studies which compare students at different institutions, find that variables such as first-generation status and admit rates can be used to predict a school's graduation rate. But when individual schools look at their own data, often those same variables have little predictive value.

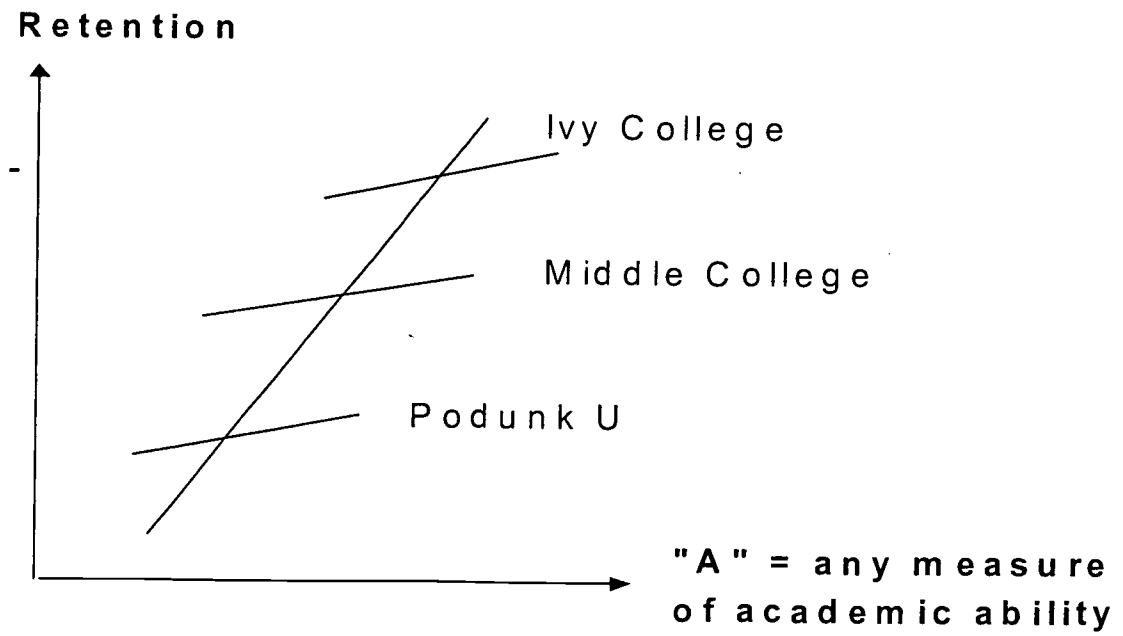
I believe that there are theoretical reasons why this pattern may frequently arise. In this paper I describe the theoretical reason, and then apply that theory to a panel of data which permits me to estimate both the inter-institutional and intra-institutional effects of schools' admit rates on their graduation rates. "Panel data" are data from a cross-section of schools – but observed longitudinally, across time.

This model, and the panel data, can be best estimated using hierarchical linear models. At yet another CAIR conference some 10 years ago, some presenters from UCLA described hierarchical linear models. Nowadays there are several statistical packages, including SAS and HLM, which can estimate hierarchical linear models with a minimum of programming expertise needed. In this paper I give a brief description of hierarchical linear models and how to estimate them using SAS.

III. A THEORETICAL EXPLANATION

Figure 1 shows possible regression lines, if we regressed any measure of retention on any variable ("A") which may affect retention – "A" could represent test scores, socio-economic status, admit rate, high school record, or any such variable. Different

FIGURE 1: A POSSIBLE EXPLANATION: NATIONAL VS INSTITUTIONAL DATA

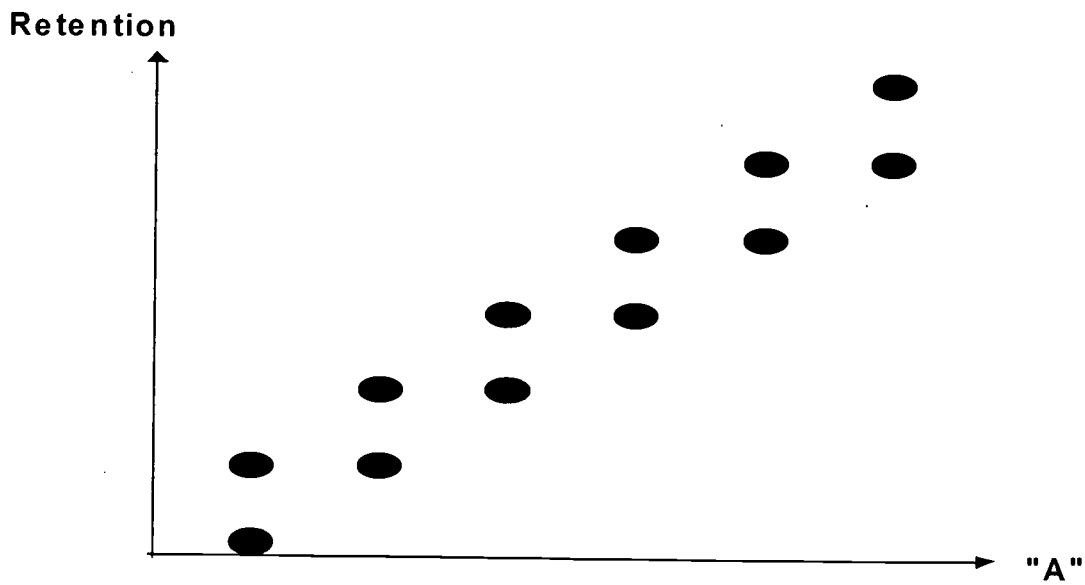


collleges have students of different academic abilities, and some of them ("Ivy College") have higher graduation rates as a result. Yet the regression line for any one college is shallower than the national regression line.

A possible reason for this pattern is that students at each school do NOT represent a random sample from the national population. Instead there are important self-selection factors involved: schools can choose to admit or not admit a student (on the basis of "A" or any other characteristics), and students can choose to matriculate at that school or not.

To see this, start with Figure 2, which shows hypothetical national data showing the relationship between "A" and retention.

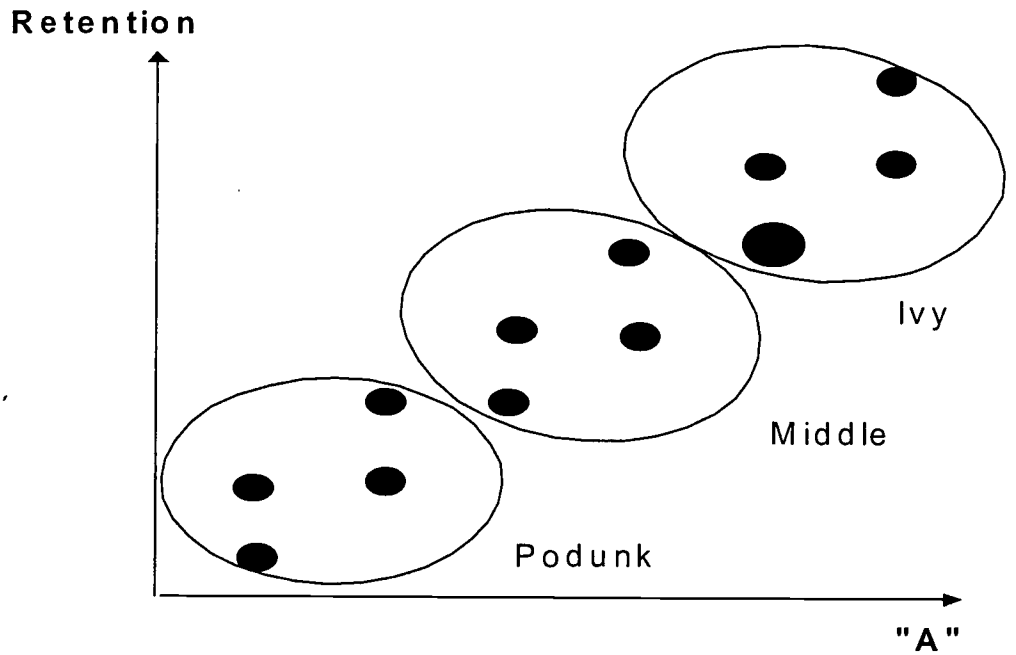
FIGURE 2: WHY WE MIGHT EXPECT TO SEE FLATTER INSITUTIONAL REGRESSION LINES (Part1 of 4)



"A" = any measure of academic ability

If students went to schools based solely on "A", for example on the basis of their SAT scores, then each school might have data that appears as in Figure 3.

FIGURE 3: IF STUDENTS WENT TO SCHOOLS BASED ON "A" (Part 2 of 4)



In Figure 3, some schools have high SAT students and high retention rates, others have lower scoring students and retention rates. But each individual school has data that show the same relationship between retention and "A" – each individual school has a regression line with the same slope.

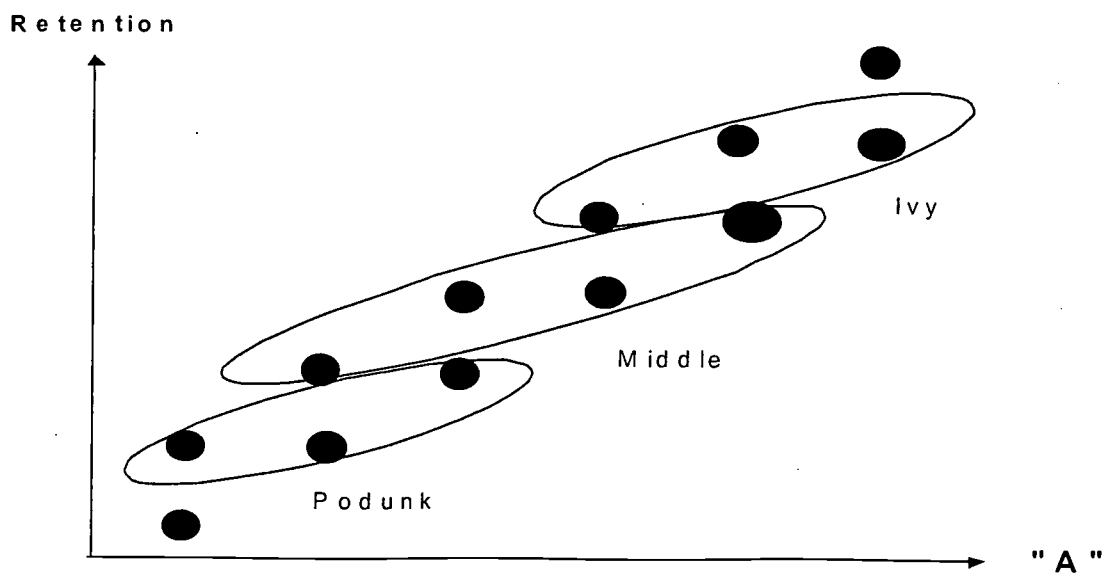
However, there is good reason to believe that many students will not go to schools solely on the basis of "A". We can hypothesize that some students may look superficially similar on the basis of "A", but end up at different schools – due to important differences, differences which are not measured by "A".

This is perhaps best illustrated by using an example: let "A" be students' combined SAT scores. Consider students with 1200 SAT scores. In general, they can be expected to have higher graduation rates than students with 1100 SATs, based on national data.

But at any one school, that relationship may be much weaker. If we observe one such 1200 SAT student at high selectivity "Ivy College", another attending "Medium College" and a third attending low selectivity "Podunk University", it is reasonable to believe that those students differ from each other in important ways. The student who got admitted into Ivy College has important characteristics, not measured by the 1200 SAT score, which (a) got him or her admitted into Ivy College and (b) will make him or her more likely to graduate. These unobserved characteristics could be high school grades, writing ability, self-discipline, or any other characteristics which lead to matriculation at a college such as Ivy College, and which also lead to higher probabilities of graduation.

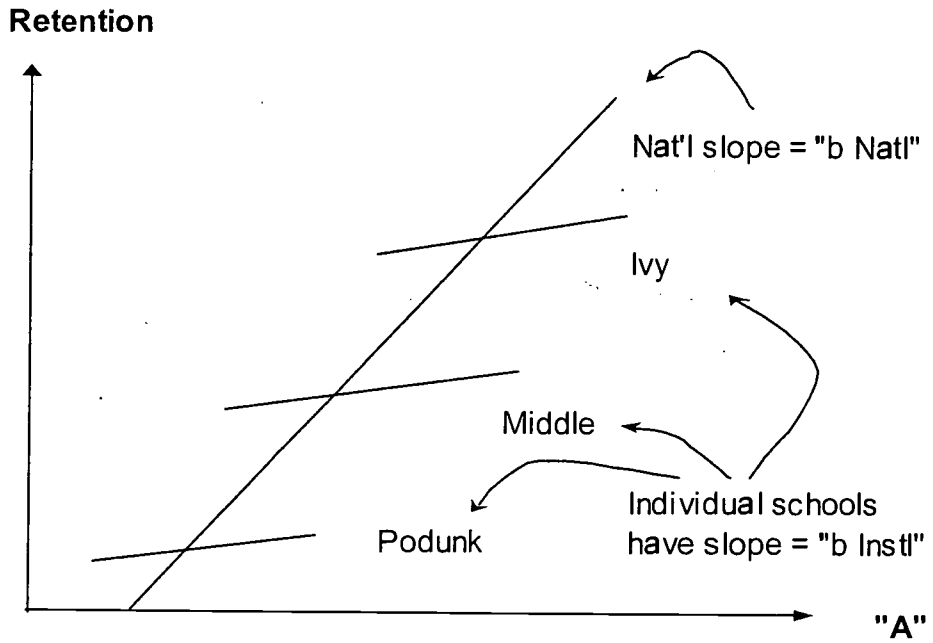
This is illustrated in Figure 4; note the flatness of the regression line for each individual school. The big dot in Figure 4 might represent a student with a 1200 SAT; in Figure 3 this person went to Ivy College but in Figure 4 she goes to Middle College. Compared to other students with 1200 SAT scores, she has a lower probability of graduating, due to characteristics which may be detectable by Ivy College.

FIGURE 4: IN FACT, SCHOOLS AND STUDENTS MAKE CHOICES -- NON-RANDOM OUTCOMES (Part 3 of 4)



So the result is regression lines that have the relationship that we've observed with the reports from Pasadena City College, the University of LaVerne, and Lawrence University: when we look at students (or schools) from across the nation, we observe a relatively steep regression line – one with slope "b Nat'l" in Figure 5. When we look at students from an individual school (or at a school across time), we observe shallower regression lines – ones with slope "b Inst'l" in Figure 5.

FIGURE 5: SO THE NATIONAL DATA AND INSTITUTIONAL DATA WILL SHOW DIFFERENT REGRESSION SLOPES (Part 4 of 4)



IV. PANEL DATA

To investigate the phenomenon of differing national and institutional relationships, we need a data set that combines national and institutional data. Because data on admit rates and graduation rates are easily available, I investigated their

relationship. There are of course many other factors which affect students' retention rates, but in this study I am not attempting to look at all possible factors. I'm focusing on one factor – admit rates – and seeing how it is related to graduation rates, at the national and the institutional level.

If we have data on admit rates and graduation rates from one year on a variety of schools, we have what are called "cross-sectional" data. But to see how differing admit rates at one school affect its graduation rates, we need to observe that school for several years – in other we need what are called "time series" data.

In other words we need "panel data": a cross-section of schools, observed across time.

I have gathered two panel data sets. The first one includes data from the Higher Education Data Sharing Consortium (HEDS) and includes only private schools. I looked at 4-year and 5-year graduation rates and admit rates. Because much of those data are proprietary, I created a second data set using only publicly available data: 6-year graduation rates from NCAA Graduation Rate reports, and admit rate data from the America's Best Colleges Guide from USNews and World Report.

The first data set includes 13 schools, observed for up to 5 years (fall 1992 through fall 1996 freshman cohorts). Results from this data set are not reported in this paper, but are very similar to the results from the second data set. The 13 schools are:

Beloit College	Brandeis University	Carleton College	Claremont McKenna
DePaul University	Grinnell College	Lawrence University	Occidental College
Reed College	Swarthmore College	Tulane Univesity	Univ of Notre Dame

Whittier College

The second data set, whose analysis will be reported in this paper, has 28 schools observed for up to 7 years (fall 1988 through fall 1994 freshman cohorts). The schools are:

Beloit College	Brandeis University	Bucknell University	Carleton College
Cleveland State Univ	CSU Fresno	DePaul University	Drake University
Grinnell College	Harvard University	Jacksonville Univ	Lafayette College
Lamar University	Lawrence University	Occidental College	Ohio State Univ
Rutgers University	San Jose State Univ	SE Lousiana Univ	Southern Utah U
Swarthmore College	Tulane Univesity	Univ of Notre Dame	Univ of the Pacific
UC Berkeley	UCLA	USC	Whittier College

V. GRAPHICAL RESULTS

Figure 6 shows a scatterplot of the 6-year graduation rates and admit rates for each cohort at each school. Notice that the scatterplot overall shows a strong relationship between graduation rates and admit rates. But not shown in this scatterplot are the identities of the individual schools. For that, see Figure 7.

FIGURE 6: 6-YEAR GRADUATION RATES AND ADMIT RATES

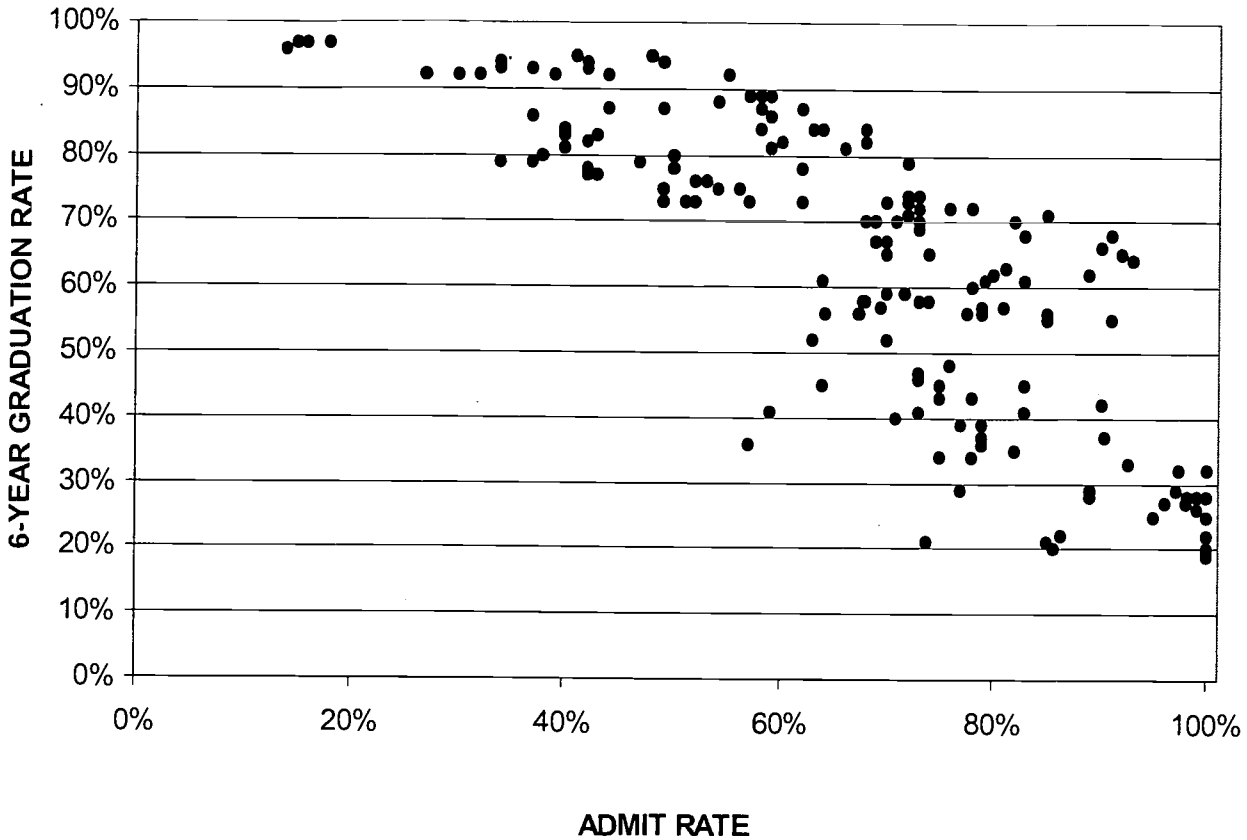
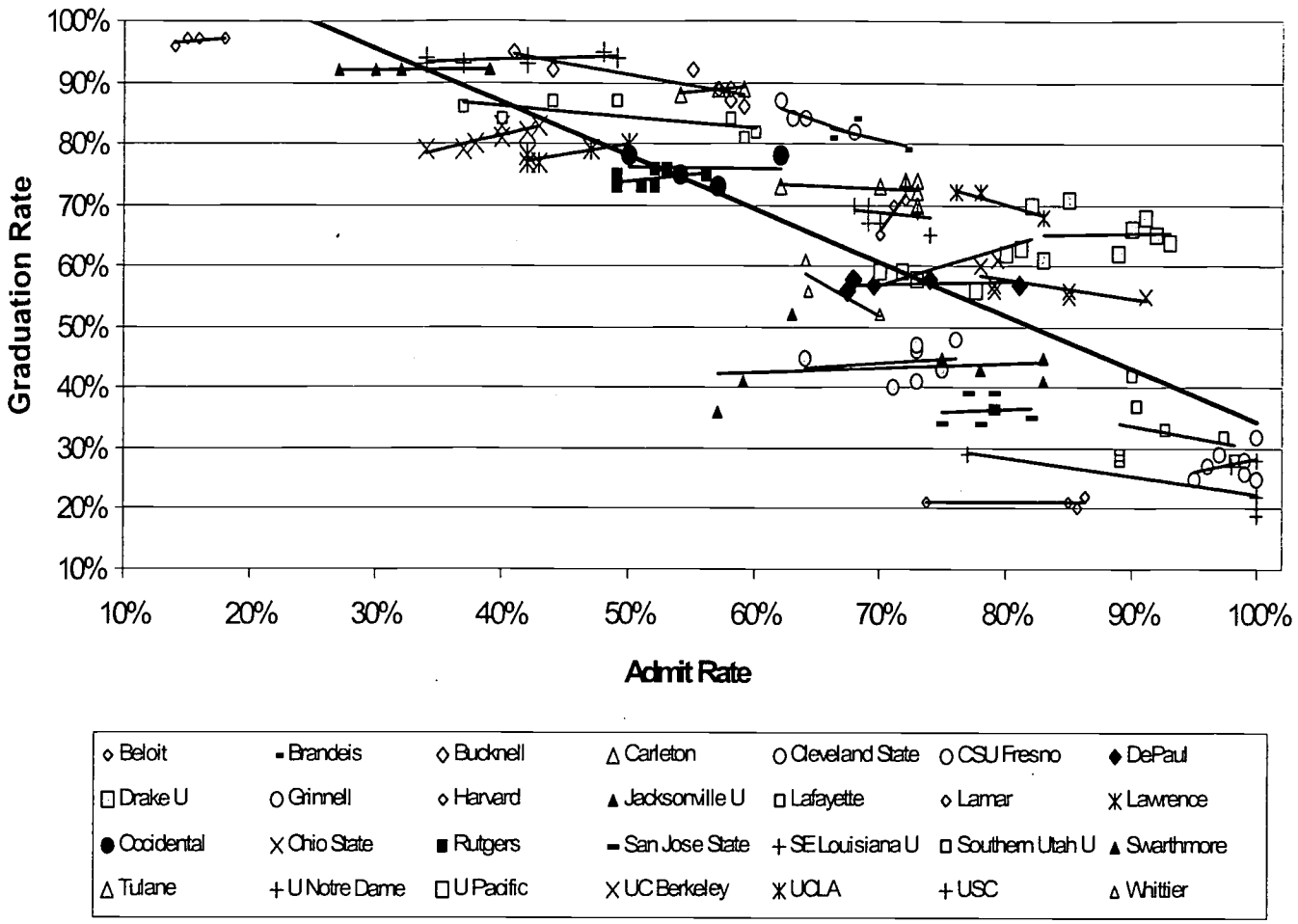


Figure 7 illustrates how the data points for individual schools cluster tightly on the scatterplot. Figure 7 also has a dark line showing a simple regression line running through the entire data set, and many lighter lines showing simple regression lines through the observations for each individual school.

FIGURE 7: 6-YEAR GRAD RATES AND ADMIT RATES, BY SCHOOL



Although some schools such as Bucknell, Grinnell, and Southeast Louisiana State exhibit strongly downward-sloping regression lines, most schools have regression lines which are appreciably shallower than the national line. Several schools, including UC Berkeley, UCLA, Beloit College, and the University of the Pacific, have lines which even seem to slope upward.

Thus this graph suggests that the data support the hypothesis illustrated in Figure 5: nationally, the relationship between schools' admit rates and their graduation rates is strong. But for any one school, variations in its admit rate often do not cause much

change in its graduation rate.

The graphs tell a good story, but to get a quantitative estimate of the slope of the national regression line ("b Nat'l") and of the slope of the typical institutional regression line ("b Inst'l"), we need a type of regression analysis (or Analysis of Covariance) which permits us to estimate those slopes in one big model. Fortunately, such models are readily available: hierarchical linear models.

VI. HIERARCHICAL LINEAR MODELS

At yet another CAIR Conference in the early 1990s, some presenters from UCLA talked about hierarchical linear models. Such models are useful when there are two or more levels of explanatory variables. For example, students' reading scores in elementary school can be partly explained by a variety of variables associated with individual students – their IQ, gender, socio-economic status, etc. But their reading scores may also be affected by the quality of the elementary school they attend, the quality of the teacher, the curriculum used, etc. – in other words, variables which are associated with an entire classroom of students, not with an individual student.

One can estimate the effects of all of these variables using standard statistical techniques such as multivariate linear regression. But such techniques do not make full use of the qualities of the data, in particular the fact that several groups of students will have shared characteristics, namely the classroom that they share. This is somewhat analogous to the difference between the graphs in Figures 6 and 7. Figure 6 failed to illustrate some important relationships in the data, such as that Harvard students attend a school with consistently low admit rates and high graduation rates (though not higher

than predicted by the national regression line).

Hierarchical linear models in this example can be thought of as estimating two levels of relationships: the relationship between individuals' characteristics and their reading scores, and the relationship between classroom's characteristics and their mean reading scores. In general, hierarchical linear models can be used whenever you have explanatory variables from different "levels".

In the case of admit rates and graduation rates, we are trying to measure the impact of admit rates on two levels: the institutional level (looking at one school across time) and the national level (looking at a number of different schools). And hierarchical linear models permit us to do this.

According to Bryk and Raudenbush (1996), hierarchical linear models are used in several fields, often under different names: "multilevel models" in sociology, "mixed-effects" and "random-effects" models in biometrics, "covariance components models" in statistics, and "random coefficient regression models" in econometrics.

In recent years, statistical packages have appeared which estimate hierarchical linear models automatically: HLM and SAS to name two. In the case of SAS, one uses the procedure Proc Mixed.

VII. SPECIFICATION OF THE MODEL

With our panel data set, the two levels of the hierarchical level model are the national level and the institutional level. At the national level, we are seeing how different schools' admit rates affect their graduation rates. At the institutional level, we are seeing how one school's admit rates affects its graduation rates.

We can't use the same explanatory variable, admit rates, in both levels. That is essentially like trying to use one variable twice as an explanatory variable in a regression. Fortunately there is an easy and natural solution: at the national level, look at each school's overall admit rate – in other words, its mean admit rate. Harvard on average has low admit rates and high graduation rates. Non-selective schools have the opposite. So at the national level, mean admit rates provide the explanatory power we need to distinguish the Harvards from the non-selective schools.

At the institutional level, an alternative to looking at a school's admit rates and graduation rates is to look at how changes in its admit rates cause changes in its graduation rate. So we will use as an explanatory variable the DIFFERENCE between the school's admit rate in any given year and its mean admit rate.

Here are the equations for this model. Let "adm_mean" be a school's mean admit rate, "adm_dif" be the difference between its admit rate and its mean admit rate, and "y" be the 6-year graduation rate for that cohort at that school. The letters "i" and "t" are indices for schools and time respectively. That is, i ranges from 1 to 28 because we have 28 schools. t ranges from 1988 to 1994 because we are looking at cohorts from those years. The letter "b" indicates a parameter to be estimated by the model, i.e. a slope coefficient or a constant term from the regression.

At the institutional level, the equation is

$$yit = bi + (bi\ Instl) * adm\_dif + eit \qquad (1)$$

where bi is the constant term (different for each school, hence the index letter i), and will

be equal to each school's mean graduation rate. "bi Instl" is the slope coefficient for adm_dif; this is the parameter which measures the relationship between changes in a school's admit rate and its graduation rate. eit is the random error term, also known as a residual.

The hypothesis is that the "bi Instl" values are relatively small for most schools, that the relationship between changes in a school's admit rates and its graduation rate is a weak one.

At the national level, the hypothesis is that there IS a relationship between a school's mean admit rate and its graduation rate. That is, the bi term will be higher for certain schools, such as Harvard, which have low adm_mean values. In addition, we want to recognize the possibility that different schools might have different "bi Instl" values (e.g. Figure 7 suggests that Bucknell has a fairly large negative "bi Instl" but UC Berkeley and UCLA may have positive ones). So the equations for the second level of the hierarchical linear model are

$$bi = c0 + (b\ Natl)*adm\_mean + ui \qquad (2)$$

$$(bi\ Instl) = b0 + vi \qquad (3)$$

The first equation says that each school's constant term, bi, is determined by a universal constant c0, the school's mean admit rate adm_mean, and a residual term, or random error term, ui.

The second equation says that each school's adm_dif slope coefficient, "bi Instl", is determined by an overall slope coefficient b0, and a residual term vi.

We can combine these equations into a single equation for the whole model

$$yit = c0 + (b\ Natl)*adm\_mean + ui + b0*adm\_dif + vi*adm\_dif + eit \quad (4)$$

There are two explanatory variables in this equation, adm_mean and adm_dif. Adm_mean is what is known as a "fixed effect" because it is in the model as an ordinary explanatory variable. Adm_dif is a "random effect" because in addition to having the fixed slope coefficient b0, it also a random effect vi*adm_dif. The constant term is also a random effect; in addition to the fixed coefficient c0, it also has a random effect ui.

These models which combine fixed effects and random effects are often called "mixed models."

There is one other aspect of model specification which we must address: linearity. The scatterplots in Figures 6 and 7 suggest that the relationship may not be a linear one; the scatterplots appear to have some concavity. A simple linear regression including a squared term for admit rates as well as a linear term confirmed this. This problem is not uncommon when the dependent (and for that matter independent) variables are measured as percentages or proportions, and are thus limited to a range from 0 to 1.

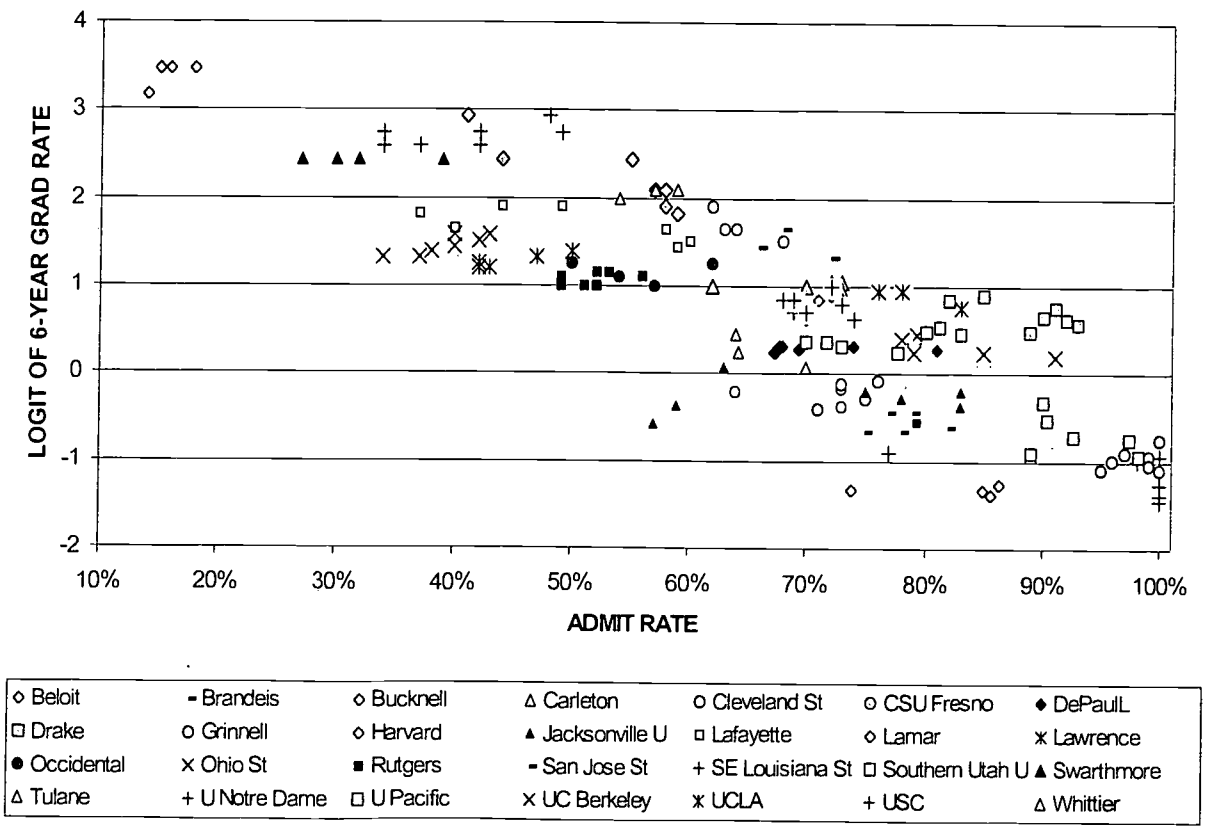
A good solution is to transform the data so their range is not limited. A commonly used transformation is the logit or logistic transformation. If "p" is a probabililty, the logit of p is

$$\text{logit} = \ln((p/(1-p))) \qquad (5)$$

p/(1-p) is the odds transformation (transforming a probability into the equivalent odds). A logit is the natural logarithm of the odds. The transformed variable has a range from negative infinity to positive infinity and thus unlike probability does not have a restricted range.

I transformed the schools' graduation rates using the logit transformation. I did not do so for their admit rates; while such a transformation might be desirable, some schools in some years had 100% admit rates, and thus the logit function cannot be applied to them (it would require dividing by 0). The transformed data are displayed in Figure 8.

FIGURE 8: LOGIT TRANSFORMATION



This scatterplot appears to show greater linearity than the plots with untransformed data, and a regression with squared admit rate had a non-significant coefficient, suggesting that there may be no significant non-linearities.

So the actual estimation of the hierarchical linear model uses the logit of schools' 6-year graduation rates as the dependent variable, yit.

Here are some basic statistics for the variables in the data set. There are 28 schools, each with up to 7 years of data (fall 1988 to fall 1994). 165 observations total.

Variable	Mean	Std Dev	Min	Max
Grad6	.64	.23	.19	.97
Logit6	.77	1.22	-1.45	3.48
Admit	.66	.21	.14	1.00
Adm_mean	.66	.20	.16	.98
Adm_dif	0.0	.05	-.19	.12

VIII. SAS PROC MIXED

In recent years a number of statistical packages have appeared which can estimate hierarchical linear models automatically. HLM is one such package, and SAS is another, with version 8's (and I believe version 7's) proc mixed procedure.

The SAS command for estimating the hierarchical linear model that we have

hypothesized is

```
Proc mixed;
        Class college;
        Model logit6 = adm_mean adm_dif /solution;
        Random int adm_dif /type=unstructured subject=college solution;
Run;
```

"Class college" simply tells SAS that college is a categorical variable, as opposed to numeric. The "model" command tells SAS that the dependent variable is logit6, and the explanatory variables are adm_mean and adm_dif, plus a constant term (included by default). The "random" command tells SAS that the constant term (intercept) and adm_dif are random effects, not fixed effects. The "/type" command tells SAS what error or residual structure to assume. "Unstructured" makes the least assumptions. The "subject=college" command tells SAS that we have panel data with "college" identifying the different "subjects." The "/solution" commands tell SAS to display individual coefficient estimates, instead of suppressing them.

IX. RESULTS.

The equation for the model again is

$$yit = c0 + (b\ Natl)*adm\_mean + ui + b0*adm\_dif + vi*adm\_dif + eit$$

The estimates for the fixed terms are

	Coeff	Std Err	t-stat
constant =	4.39	.394	11.15
b Natl =	**-5.44**	.57	**-9.51**
b Instl =	**-.40**	.42	**-.94**

As hypothesized, adm_mean has a strong relationship with a schools' graduation rate, with a highly significant ($p<.0001$) t-statistic. However, overall adm_dif does not ($p=.36$); changes in a school's admit rates over time do not show a relationship with its graduation rate, on average.

The random terms (ui, vi, and eit) have the following estimated variances:

Estimated variance of ui = .38

vi = 2.42

eit = .018

The individual schools have the following estimated vi terms (recall from equation (3) that their slope coefficients on adm_dif are the sum of "b Instl" and vi):

College	Estimate of vi	Std, Error	T-value	Pr > |t|
Beloit	-0.01	1.37	-0.01	0.995
Brandeis	-1.21	1.27	-0.95	0.342
Bucknell	-3.50	0.74	-4.74	<.001
Carleton	-0.49	1.31	-0.38	0.708
Cleveland State	0.51	1.26	0.41	0.686
CSU Fresno	0.83	1.01	0.82	0.413
DePaul	0.42	0.90	0.47	0.638

Drake U	-0.46	1.05	-0.44	0.660
Grinnell	-1.60	1.26	-1.27	0.208
Harvard	0.71	1.34	0.53	0.599
Jacksonville U	0.75	0.60	1.25	0.215
Lafayette	-0.73	0.65	-1.13	0.261
Lamar	0.90	0.99	0.91	0.367
Lawrence	-1.26	1.24	-1.01	0.313
Occidental	0.28	1.06	0.26	0.794
Ohio State	-0.76	0.93	-0.82	0.412
Rutgers	0.85	1.20	0.71	0.480
San Jose State	0.79	1.23	0.64	0.523
SE Louisiana U	-0.97	0.69	-1.41	0.162
Southern Utah U	-0.52	1.02	-0.51	0.614
Swarthmore	0.31	1.06	0.29	0.772
Tulane	-0.24	1.00	-0.24	0.812
U Notre Dame	1.00	0.82	1.22	0.225
U Pacific	1.75	0.92	1.9	0.060
UC Berkeley	1.91	1.12	1.71	0.090
UCLA	1.44	1.11	1.3	0.198
USC	-0.38	1.21	-0.31	0.755
Whittier	-0.31	1.25	-0.25	0.806

The individual schools have the following estimated ui terms (recall from equation (2) that each school's constant term or intercept is the sum of the constant c0 and ui):

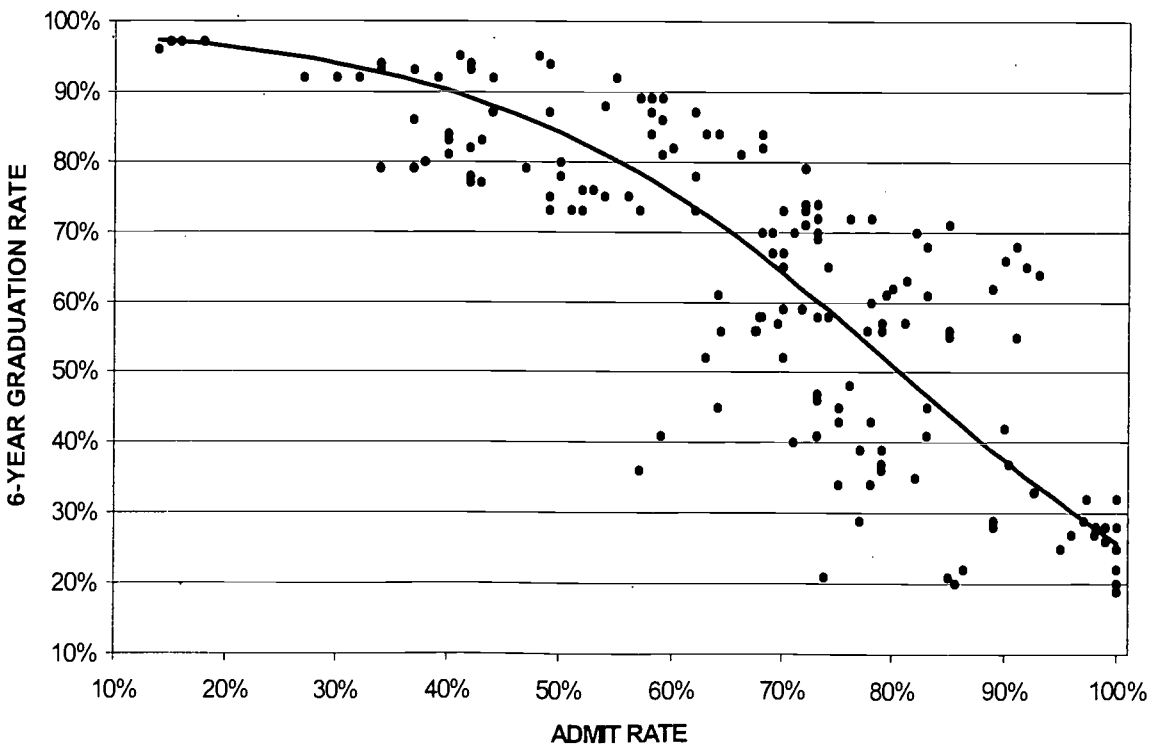
College	Estimate	Std, Error	T-value	Pr > |t|
Beloit	0.26	0.14	1.80	0.075
Brandeis	0.81	0.14	5.80	<.001
Bucknell	0.75	0.15	5.14	<.001
Carleton	0.75	0.14	5.27	<.001
Cleveland State	-0.03	0.22	-0.15	0.881
CSU Fresno	-0.69	0.13	-5.21	<.001
DePaul	-0.25	0.13	-1.93	0.057
Drake U	1.08	0.18	5.88	<.001
Grinnell	0.78	0.13	5.83	<.001
Harvard	-0.08	0.31	-0.25	0.804
Jacksonville U	-0.78	0.13	-6.00	<.001
Lafayette	0.01	0.16	0.06	0.951
Lamar	-1.20	0.16	-7.27	<.001
Lawrence	0.78	0.16	4.93	<.001
Occidental	-0.20	0.15	-1.36	0.178
Ohio State	0.38	0.16	2.38	0.019
Rutgers	-0.50	0.15	-3.35	0.001

San Jose State	-0.68	0.15	-4.66	<.001
SE Louisiana U	-0.34	0.22	-1.56	0.122
Southern Utah U	-0.09	0.20	-0.46	0.644
Swarthmore	-0.20	0.23	-0.87	0.386
Tulane	0.44	0.13	3.35	0.001
U Notre Dame	0.54	0.19	2.83	0.006
U Pacific	0.22	0.14	1.56	0.121
UC Berkeley	-0.80	0.20	-4.06	<.001
UCLA	-0.67	0.17	-3.85	<.001
USC	0.25	0.13	1.89	0.061
Whittier	-0.52	0.14	-3.77	<.001

X. INTERPRETATION

The most important result is confirmation of the hypothesis that "b Natl" shows a large and statistically significant negative relationship between schools' mean admit rates and their graduation rates, while "b Instl" shows no significant relationship. Figure 9

FIGURE 9: THE NATIONAL REGRESSION LINE ON MEAN ADMIT RATES SHOWS A GOOD FIT

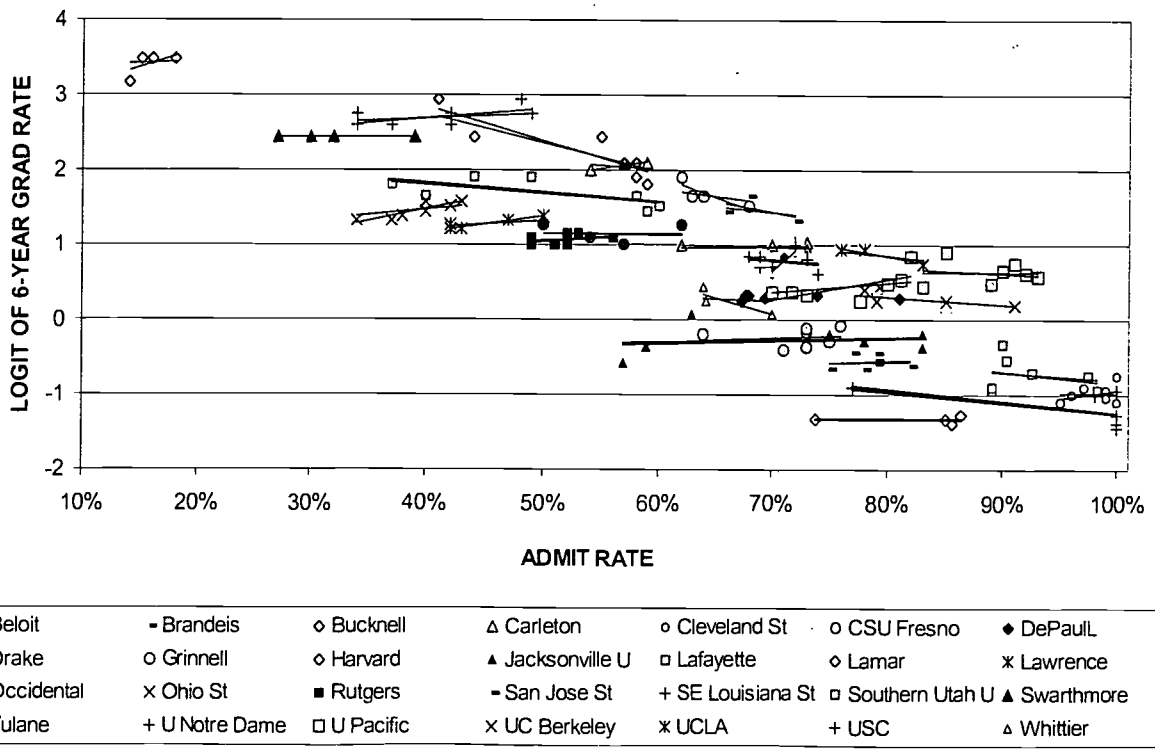


shows that national regression line, on the original scatterplot (with admit rate rather than logit6 on the vertical axis). Note that the logistic transformation, leading to a curved regression line, seems to model the data more effectively than a linear regression line.

The individual schools of course do show some diversity. Bucknell's estimated vi is both highly negative and significant, suggesting that over the 7 years covered, changes in its admit rates did indeed affect its graduation rates. But no other school had a vi whose estimate was significantly different from 0 at the p=.05 level (of course, with only 7 observations of each school, it is hard to achieve statistical significance). In short, for most schools, there is little evidence that changes in their admit rates during this 7 year period led to changes in their graduation rates.

Figure 10 (which reverts to putting logit6 on the vertical axis, for the sake of

FIGURE 10: IN MOST CASES, THE SCHOOLS' REGRESSION LINES ARE VERY CLOSE TO SIMPLE TRENDLINES



linearity) shows both the simple trendlines and the regression lines from the hierarchical linear model. We can see that in almost all cases they are very similar, so the hierarchical linear model is producing coefficients which look "correct". There are some exceptions (Beloit, Whittier, and Harvard) but a possible explanation is that Beloit and Whittier both have only three years of data, hence their estimated "b Instl" values may reflect more of the overall mean and less of their data. Harvard conceivably might be a type of outlier; despite lying close to the national regression line, its admit rates and graduation rates are outside the range of any other schools' in this sample.

The constant terms provide some hints about schools graduation rates; some schools have graduation rates which put them above the national regression line and some are below. There is a strong pattern for the schools that are most above the regression line to be private while the schools that are most below the regression line tend to be public. This is consistent with one of Astin's findings (1993). However, this is not a full-fledged model for predicting graduation rates (it uses essentially only one explanatory variable, admit rates, and ignores other predictors) so these results are only hints, not conclusions.

X. CONCLUSION.

There seems to be good graphical and statistical evidence to support the claim that at least some predictors of graduation rates, such as admit rates, have better predictive value at the national, or cross-sectional, level than at the institutional, or time-series, level. However this still leaves us with the fundamental question: do changes in admit

rates change a school's graduation rates, or don't they? If a school reduces its admit rate from 80% to 40%, will its graduation rate rise?

One possible answer is: not in the short run (7 years or less), but maybe in the long run. One or a few cohorts brought in with low admit rates will not affect a school's graduation rate much, based on the "b Instl" estimates (unless the school is Bucknell). If however a school PERMANENTLY reduces its admit rate, so that instead of being among the 80% admit rate schools it is up among the 40% admit rate schools, then maybe it would enjoy a higher graduation rate.

There may however be a better answer, one which utilizes the concept of unobserved characteristics or unobserved variables. The answer could be along these lines: we observe that schools' graduation rates differ, for reasons which are not well explained by adm_dif, i.e. by changes in their admit rates. The true factors behind schools' differing graduation rates may be unobservable (attention paid to students, individual students' inherent persistence, etc.). Thus we can only observe that some schools have higher graduation rates, and some lower, without knowing the true reason why.

But in addition to these unobservable characteristics, there are some observable characteristics of schools – such as their mean admit rates, adm_mean – which, if not actual determinants of graduation rates, are correlated with such determinants. Hence, adm_mean has a large and significant correlation with graduation rates, whereas adm_dif does not.

So a more complete investigation of schools' graduation rates would look at other variables in addition to adm_mean. Undoubtedly, the predictive value of adm_mean

would decline as these other variables were added to the model. But I believe that the pattern of stronger national correlations and weaker institutional correlations would persist.

It is also worth noting that the additional variables that we could look at include both institutional characteristics (many have already been identified by investigators such as Astin (1993) – teaching vs research orientation, Catholic denomination, etc.) and characteristics of individual students (SAT scores, gender, ethnicity, intensity and quality of academic curriculum in high school as suggested by Adelman (1999). etc.). When we start looking at individual students' characteristics, we have introduced a third level of explanatory variables, to add to the ones at the time series or cohort level (adm_dif) and to the ones at the cross-institutional level (adm_mean). Hierarchical linear models would be even more useful for these more detailed investigations.

Finally, I repeat the message of the utility of good conferences. My introduction to hierarchical linear models came at a CAIR conference, and the realization that national predictors may not work as well at the institutional level came from sharing findings at another CAIR conference. The final catalyst came from a HEDS conference. So keep on conferencing!

XI. FURTHER READING.

Singer (1998) provides an excellent introduction to hierarchical linear models, especially for SAS users. The book Littell et al (1996) is not quite as good and is more focused on SAS users. Sullivan et al do not orient their discussion to SAS, but their exposition is not as clear as Singer's. The book by Bryk and Raudenbush (1992) goes

into much more depth and gives more theoretical background. Hsiao (1986) does not focus on hierarchical linear models per se but gives an even more technical background.

REFERENCES

Adelman, Clifford (1999). Answers in the Tool Box: Academic Intensity, Attendance Patterns, and Bachelor's Degree Attainment. US Department of Education.

Astin, Alexander (1993). What Matters in College?: Four Critical Years Revisited. San Francisco: Jossey Bass.

Bryk, Anthony and Raudenbush, Stephen (1992). Hierarchical Linear Models: Applications and Data Analysis Methods. (Advanced Quantitative Methods in the Social Sciences, Number 1.) Newbury Park: Sage Publications, Inc.

Hsiao, Cheng (1986). Analysis of Panel Data. (Econometric Society Monographs, Number 11.) New York: Cambridge University Press.

Littell, Ramon C., Milliken George A., Stroup, WalterW., and Wolfinger, Russell D. (1996). SAS System for Mixed Models. Cary: SAS Institute.

Singer, Judith (1998). Using SAS Proc MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. Journal of Educational and Behavioral Statistics, 24, 323-355.

Sullivan, Lisa M., Dukes, Kimberly A., and Losina, Elena (1999). Tutorial in Biostatistics: An Introduction to Hierarchical Linear Modelling. Statistics in Medicine, 18, 855-888.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").